DOCUMENT RESUME

ED 191 880 TM 800 503

AUTHOR Forster, Fred; Karr, Chad
TITLE Using the Rasch Model to Increase the Power of Item Analysis.
PUB DATE Apr 80
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; *Item Analysis; Item Banks; Language Skills; *Latent Trait Theory; *Test Items; *Test Reliability
IDENTIFIERS Mean Square Fit; Point Biserial Correlation; *Rasch Model

ABSTRACT

A method is described which used the Rasch model for combining information across administrations of the same items in different tests to different student populations. The method was used in conjunction with the composite point biserial and mean square fit to provide more valid information about item quality. Over 700 items were administered to groups of students in grades 5 through 9. Each group was comprised of 150 to 250 students, and was located in one of three school districts. Using the level of the achievement group relative to the equal interval Rasch item calibration made it possible to superimpose all the administrations onto a single item characteristic curve (ICC). Findings included: (1) neither the point biserial nor the mean square fit was a reliable indicator of item quality, and neither appeared to be a substitute for the ICC; (2) the method of combining information across administrations significantly increased the amount and utility of information available about the actual performance of an item. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Karr

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED191880

USING THE RASCH MODEL TO INCREASE THE POWER OF ITEM ANALYSIS

Fred Forster
Portland Public Schools

Chad Karr
Portland State University

Portland, Oregon

Paper presented at Joint Session 15.05: Advances in the Application of the Rasch Model to Practical School Problems; Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, Boston, April 7-11, 1980.

The commonly used sample dependent indices of item quality such as the point-biserial correlation coefficient and the mean square fit have inherent deficiencies. These indices can produce contradictory or misleading indications of item quality. A useful adjunct to these indices is the item characteristic curve (ICC) produced by a Rasch item analysis. Because of the equal interval characteristics of cross-validated Rasch item calibrations derived from linked tests, it is practical to combine ICCs for a given item administered in different tests to different groups to obtain a composite curve which includes a wider range of achievement. The empirically obtained curve can then be compared to the theoretical logistic curve to determine various aspects of item quality.

TM 800503

Building accuracy into mental measurement is a difficult task. We use the concepts of validity and reliability to evaluate the quality of a measurement system which has both constant and variable error. The validity of an achievement test item is established by professional estimate of the correspondence between the content of the item and the instructional outcome it represents. For this reason, the content validity of an achievement test item is determined by the properties of the item built in by the test writer. The unreliability of the item is due to the properties which have been unintentionally built into it.

The reliability of an achievement test item depends upon the degree to which it unambiguously presents a task to the student. One method of assessing the reliability of an item is to compute the correlation between answers to the item and total scores on the test. When a computer is used, the favored statistic is the point-biserial correlation coefficient.

To compensate for the biasing effect of lower or higher proportions of students passing the item (p-value), one can multiply the point-biserial correlation by the standard deviation of the item $\left(\sqrt{p\ (p-1)}\right)$ to obtain an item reliability index.

Another method of assessing reliability makes use of the latent trait model which enables one to compare the actual performance of students with a theoretically ideal distribution of performance. One can calculate a measure of the correspondence between these two distributions. The most commonly used measure is the mean square fit.

Unfortunately, both of these methods (and other statistics similarly derived) have serious drawbacks, one of which is they are sample dependent. In the case of the point biserial, restriction in the range of test scores reduces its sensitivity. This impairment presents special problems in determining the reliability of items in tests that are specially designed with shorter, more concentrated ranges for use in functional level testing. Although the mean square fit is less biased by departures from normality of test range than the point biserial, the mean square fit is also an inadequate indicator of item reliability. First, it is sensitive to differences in sample size--generally low for small samples, generally high for large samples. Second, it disproportionately weights differences near the top or bottom of the item curve (based on the term $1/pq$). Finally, it does not provide an adequate indication of where the actual and ideal distributions do not fit; e.g., could the difference be ascribed to aberration in the performance of one or two students, or does it truly reflect a serious flaw in the measurement quality of the item?

There is an approach to item reliability which helps overcome some of the shortcomings of the point biserial and mean square fit. The item characteristic curve (ICC) is a graphic display of the actual performance of students which can be compared to a theoretically ideal curve (see Figure 1). The ICC shown in Figure 1 plots the relationship between the achievement score for different groups of students and the percentage of each group passing that item. A theoretical performance curve is also included. The X-axis is the achievement level assigned to the various raw scores on the test by the Rasch procedure (i.e., each raw score corresponds to a specific achievement level). The Y-axis records the percent of students at a specific achievement level who answered the item correctly. Therefore, if 10 students had a raw score of 15 corresponding to an achievement level of 190 and if 4 of these students answered the item correctly, an "X" would be placed corresponding to 190, 40%.

A convention followed in our work is to plot the curve predicted by the Rasch equation as a series of dots and to plot the actual performance of students as an "X" or number corresponding to the score group. We use the letters A to Z to represent score groups with 10 to 35 students.

The power of the plot of the ICC is its ability to summarize the full set of information concerning the performance of students on a given question. Although some minimal training is required to interpret the discrepancies revealed by the plot, it often provides a valuable guide for detecting defects in an item. For example, if the actual curve is above the theoretical curve at the low end of the achievement scale, students may be using an incorrect method which leads to the correct answer (see Figure 2-A). On the other hand, if the actual curve drops off at the high end, there may be a subtle problem which tends to entrap some of the most able students (see Figure 2-B).

Although the item curve plot overcomes most of the drawbacks inherent in the statistical approaches, it does not always provide the information needed to establish the reliability of the item. For example, if an item is too easy for a group, the information on the plot will be limited to the top portion of the curve (see Figure 2-C). For this reason, it was highly desirable to develop a method for pulling together information across separate administrations of the item to students at different achievement levels. For example, by combining the information across the three limited administrations represented in Figure 3, it is possible to arrive at the composite plot shown in Figure 4, reflecting the performance of the item across the full range of student achievement. The problem addressed by this study was the development of a procedure for combining information on item reliability across administrations of the items to different groups of students.

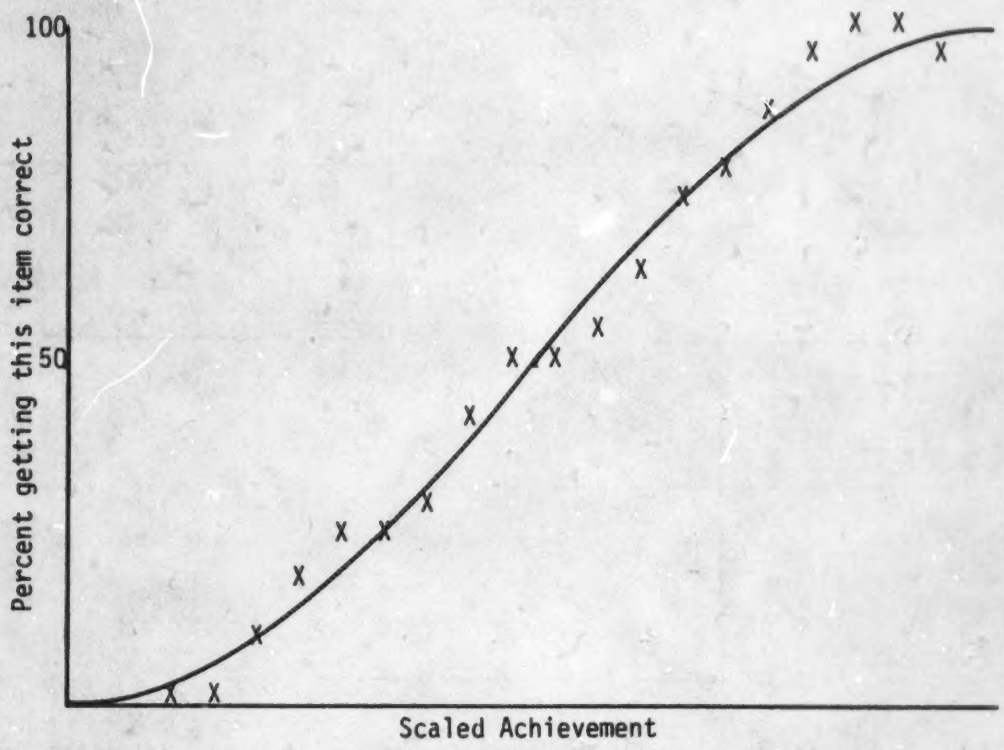


Figure 1. An item characteristic curve.

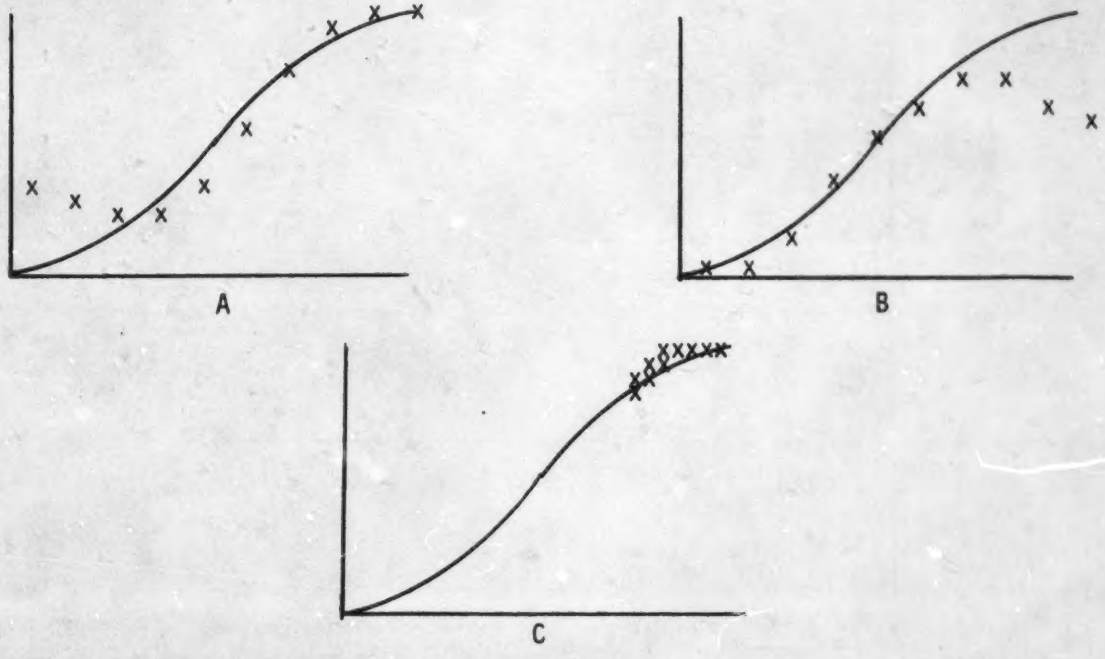


Figure 2. Examples of item characteristic curves.

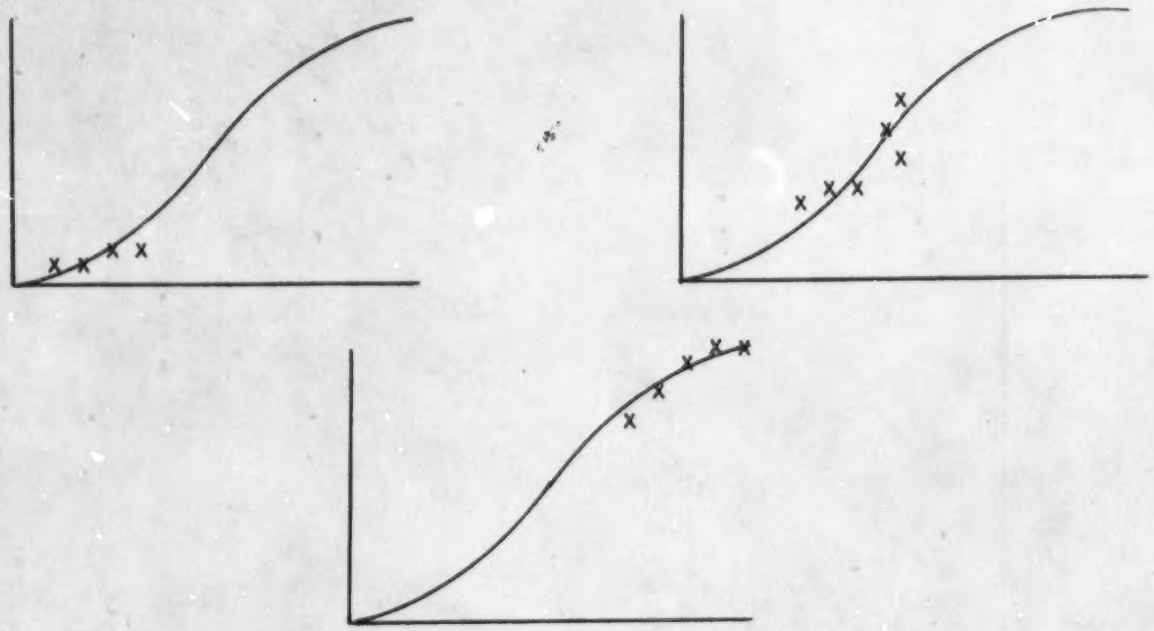


Figure 3. Item characteristic curves for three separate administrations of the same item to groups with different achievement levels.

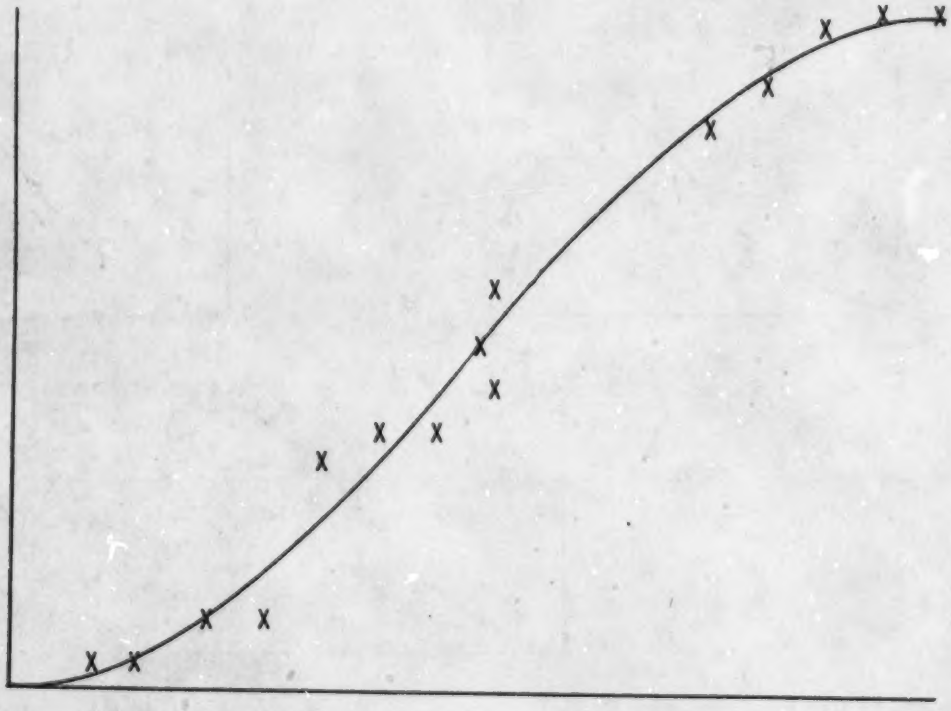


Figure 4. The composite item characteristic curve for the three separate administrations shown in Figure 3.

METHODOLOGY

Overview

This study was part of a larger project intended to build a Rasch calibrated item bank in the subject area of language usage (punctuation, grammar, spelling, etc.). A total of 128 overlapping test forms were developed to provide three independent calibrations for each item and five cross validations of the linking values between each pair of tests (see Appendix A).

Population

Three school districts participated in this project: (a) a rural Michigan district serving a predominantly white middle-class population; (b) an urban California district with a broad range of ethnic and socio-economic diversity; and (c) an urban Oregon district. The students were in grades 5 through 9.

Procedure

Each test form was administered to between 150 / 250 students in one of the three districts. The test results were Rasch analyzed, calibrations were calculated for each item, and a file was produced summarizing the performance of each raw score on each item (i.e., all students with a raw score of 1, a raw score of 2, etc.).

Since each item was responded to by three different groups of students taking different combinations of items, we developed a method for combining the information across administrations. Using the level of the achievement group relative to the equal interval Rasch item calibration made it possible to superimpose all the administrations onto a single ICC plot. For example, if an item had a calibration of 210 on one administration and 195 on a second, the raw score corresponding to 200 on the first administration would be plotted at the same point on the X-axis as that raw score corresponding to 185 on the second administration. The program we use calculates the composite point biserial, mean square fit, and ICC, using the achievement level of the students relative to the calibration of the item on a given administration.

RESULTS

The following ICC graphs show the actual results obtained by combining information across three separate administrations for a multiple-choice item included in three different forms of a language arts test. Figure 5 shows the composite ICC for an item whose performance is very satisfactory. The actual curve is steeper than the theoretical curve, which indicates a sharper discrimination than the theoretical curve. Although the item had four alternatives, there is no evidence that guessing or random selection of alternatives favored the lower achieving score groups.

Figure 6 shows another composite ICC and the three separate ICCs which are combined to make up the composite. Ninety-three percent of the students answered this item correctly. The mean square fit value of .28 is very low, indicating a very good fit to the theoretical curve. Although the point biserial is relatively high (.42), the item-reliability index is low (.11). The composite graph shows the clustering at the high end of

the curve which makes this item a poor discriminator for the full set of students to whom it was administered. The item's performance was a little better for the group in which 88% answered the item correctly. For this less high scoring group the point biserial is .53 and item-reliability index is .17.

Figure 7 shows the composite ICC for an item which varies too much from the theoretical curve in the middle range of performance. This figure also includes the three ICCs which make up the composite ICC. The mean square fit of 2.35 indicates that this item is not performing properly, but none of the indices provides as clear a picture of the problem as the ICC.

The deficiencies of the items shown in Figures 6 and 7 which led to their poor statistics and item curve characteristics are apparent when one examines the items themselves. The item in Figure 6 is clearly not well designed to test the learning objective it represents. The item in Figure 7 has distractors which are extremely different in their reading level requirements. Not all items whose flaws appear in the ICC are as apparent, but plausible reasons can usually be found on close study of such items.

Figure 5. An ICC for an item which was appropriate for the groups to which it was administered.

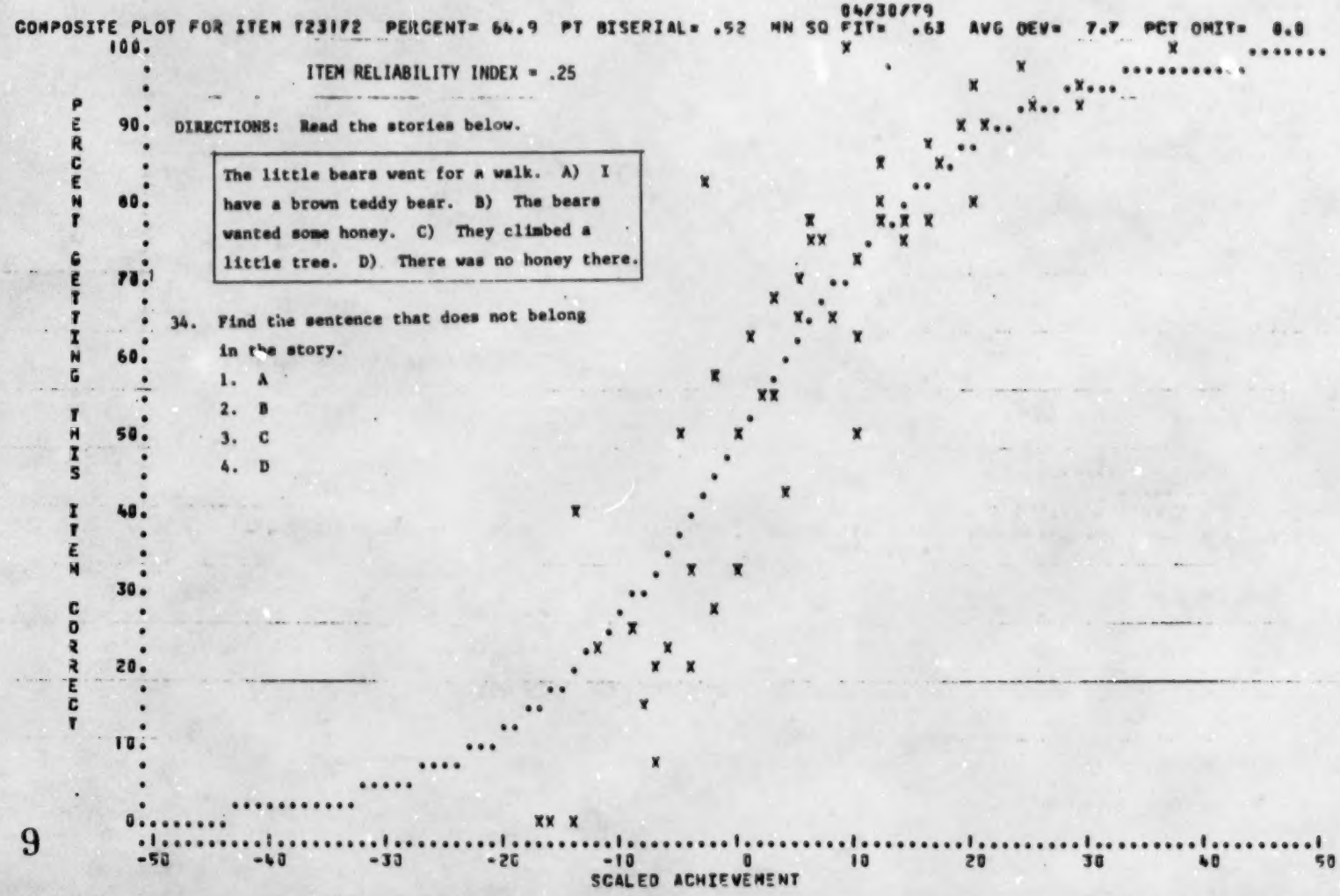

COMPOSITE PLOT FOR ITEM T23172 PERCENT= 64.9 PT BISERIAL= .52 MN SQ FIT= .63 AVG DEV= 7.7 PCT OMIT= 8.8
04/30/79
ITEM RELIABILITY INDEX = .25
DIRECTIONS: Read the stories below.
The little bears went for a walk. A) I have a brown teddy bear. B) The bears wanted some honey. C) They climbed a little tree. D) There was no honey there.
34. Find the sentence that does not belong in the story.
1. A
2. B
3. C
4. D
PERCENT GETTING THIS ITEM CORRECT
100.
90.
80.
70.
60.
50.
40.
30.
20.
10.
0.
-50 -40 -30 -20 -10 0 10 20 30 40 50
SCALED ACHIEVEMENT

Figure 6. An ICC for an item which was too easy for the groups to which it was administered.

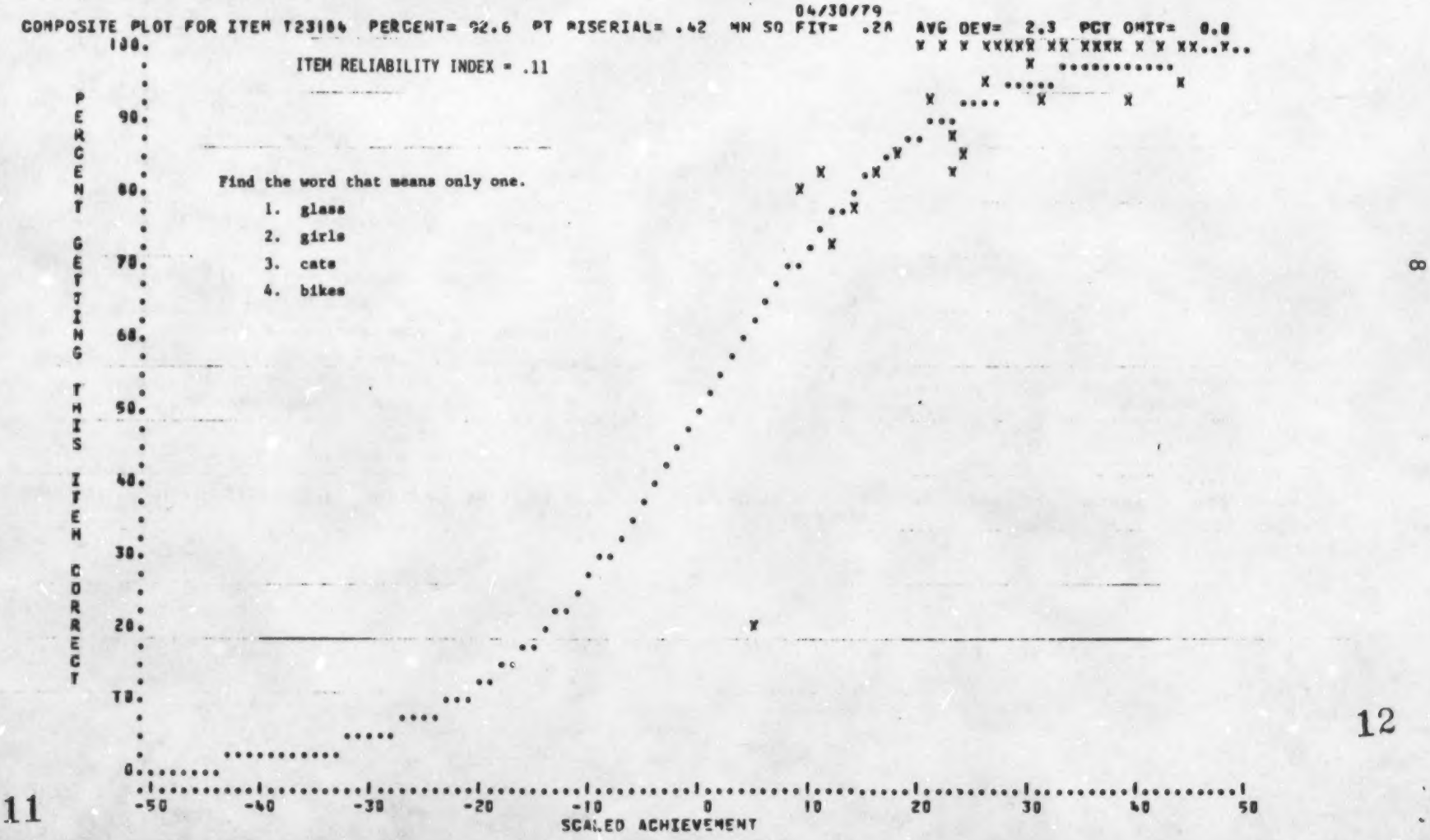


Figure 6a

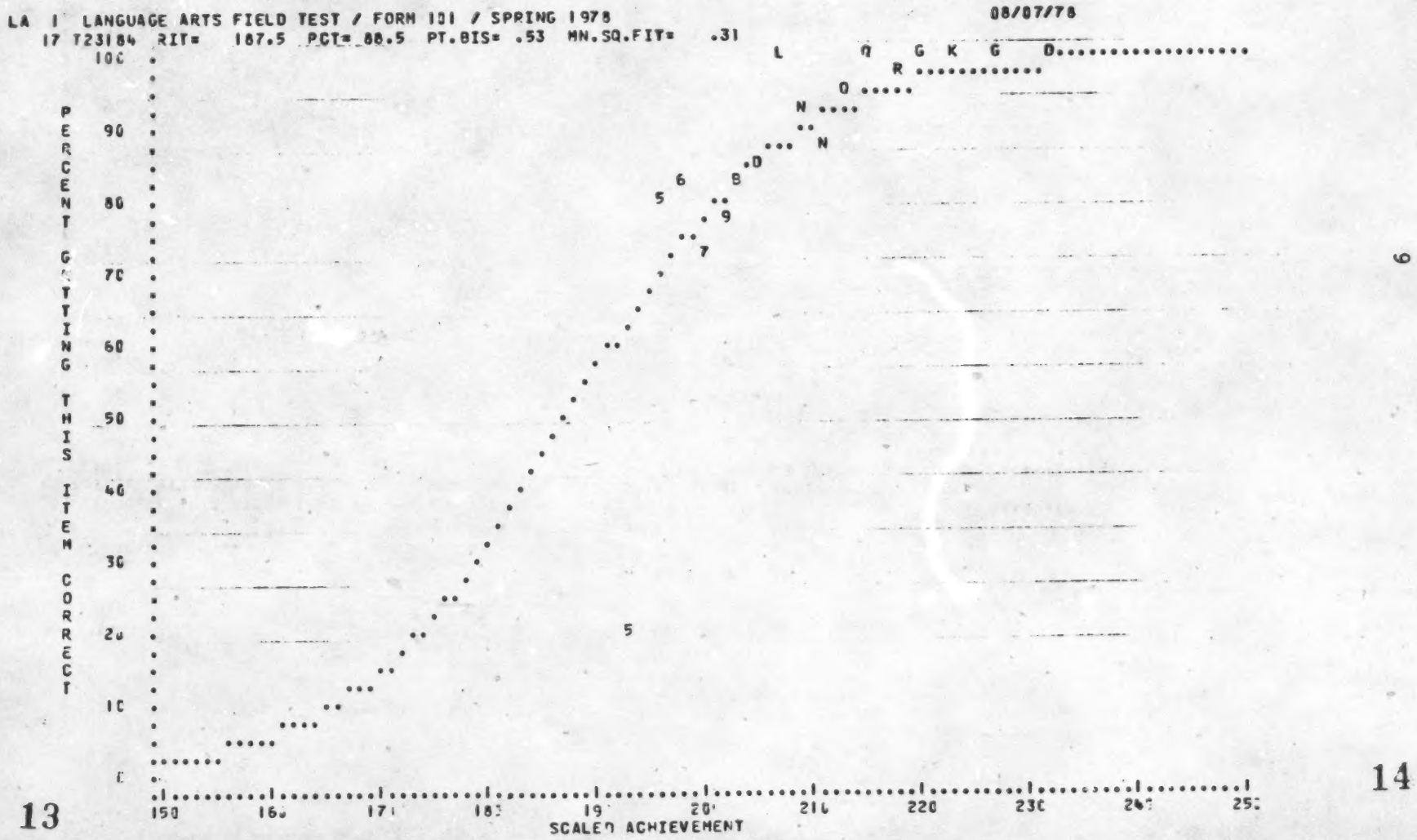

LA I LANGUAGE ARTS FIELD TEST / FORM 101 / SPRING 1978
08/07/78
17 723184 RIT= 187.5 PCT= 88.5 PT.BIS= .53 MN.SQ.FIT= .31
PERCENT GETTING THIS ITEM CORRECT
100
90
80
70
60
50
40
30
20
10
0
150
160
170
180
190
200
210
220
230
240
250
SCALED ACHIEVEMENT

Figure 6b

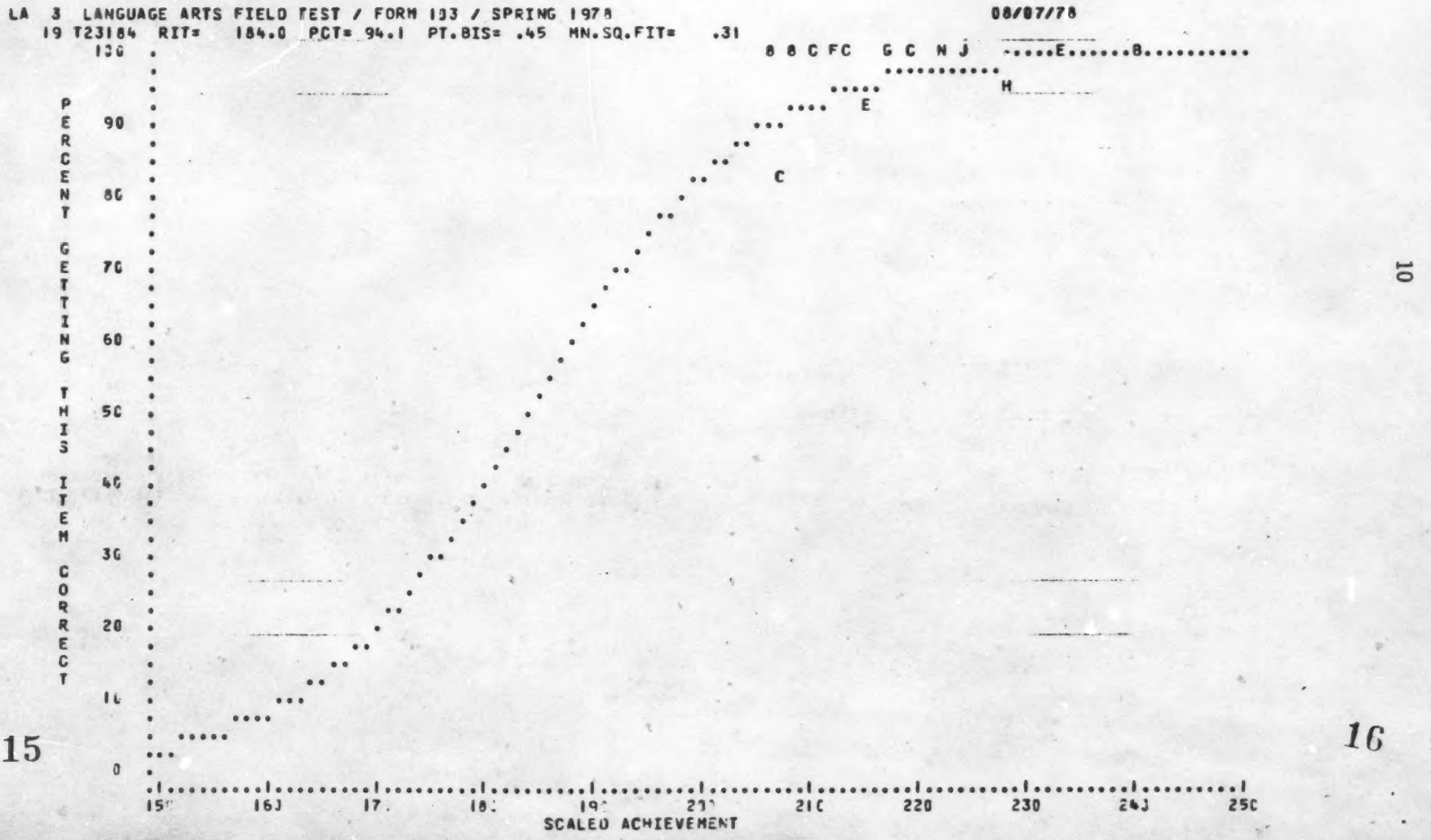

LA 3 LANGUAGE ARTS FIELD TEST / FORM 103 / SPRING 1978
08/07/78
19 T23184 RIT= 184.0 PCT= 94.1 PT.BIS= .45 MN.SQ.FIT= .31
PERCENT GETTING THIS ITEM CORRECT
SCALED ACHIEVEMENT

Figure 6c

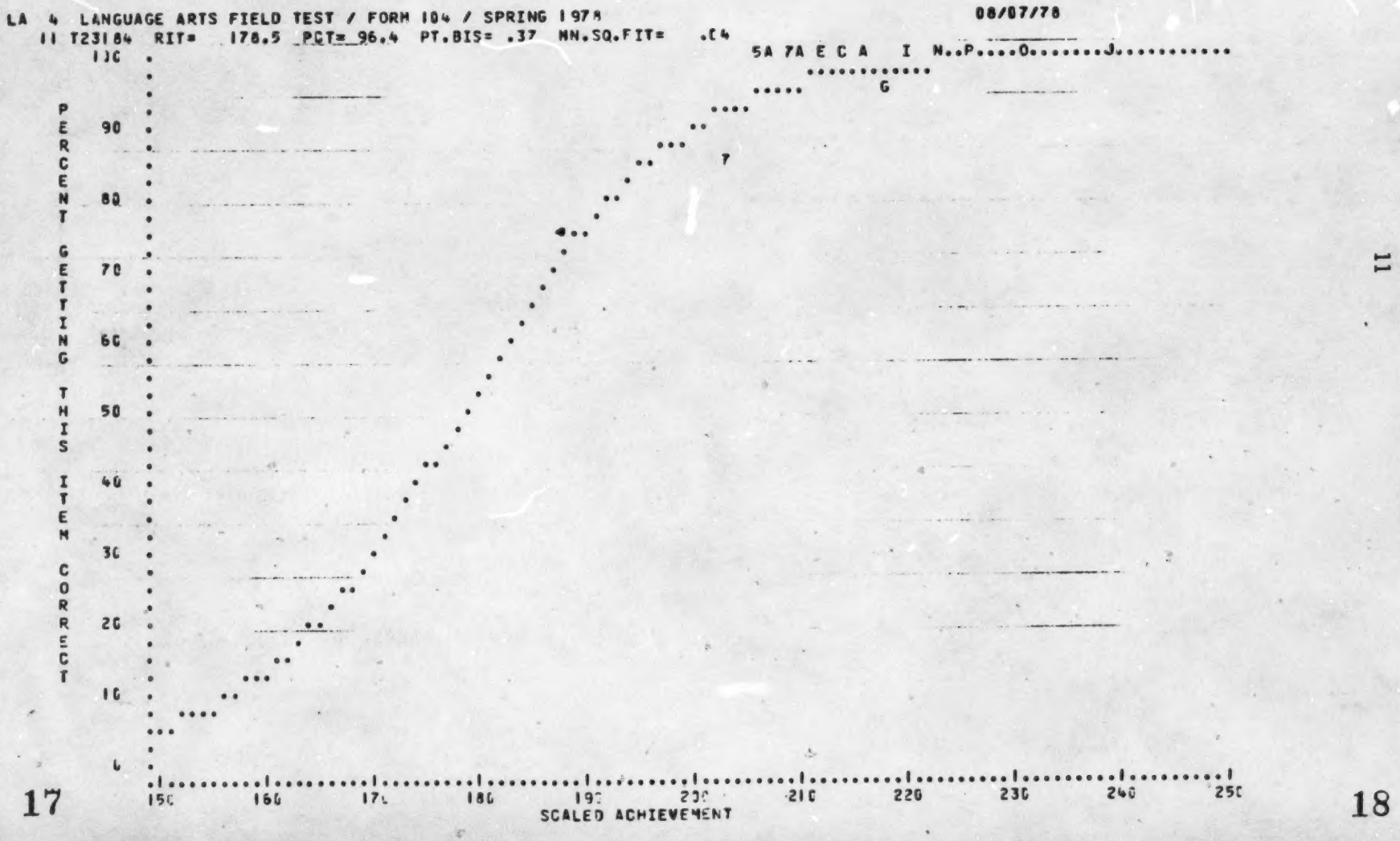

LA 4 LANGUAGE ARTS FIELD TEST / FORM 104 / SPRING 1978
08/07/78
11 T23184 RIT= 178.5 PCT= 96.4 PT.BIS= .37 MN.SQ.FIT= .04
5A 7A E C A I N..P....O.....J.....
G
PERCENT GETTING THIS ITEM CORRECT
100
90
80
70
60
50
40
30
20
10
0
150
160
170
180
190
200
210
220
230
240
250
SCALED ACHIEVEMENT

Figure 7. An ICC for an item which appears to be defective.

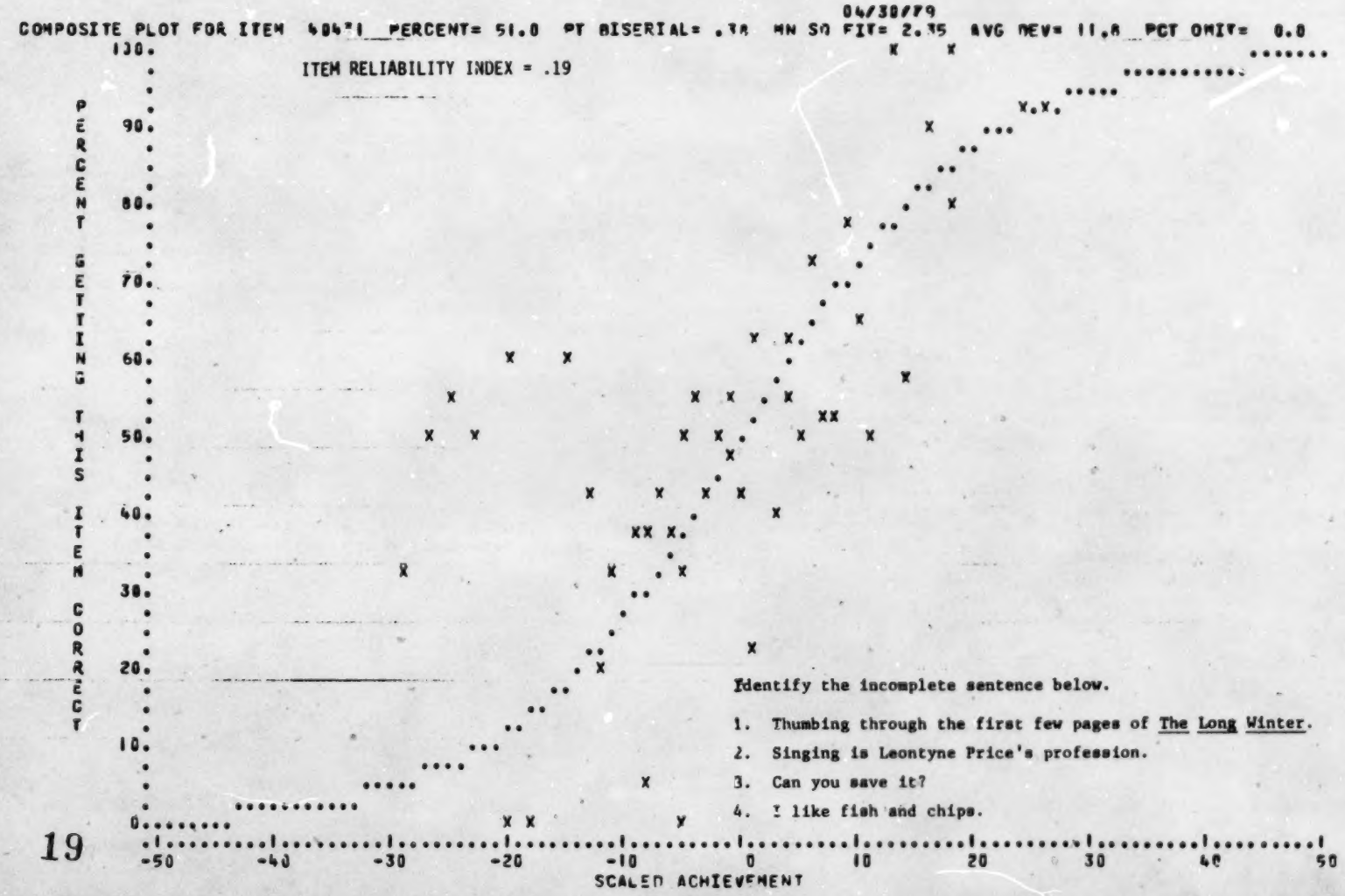


Figure 7a

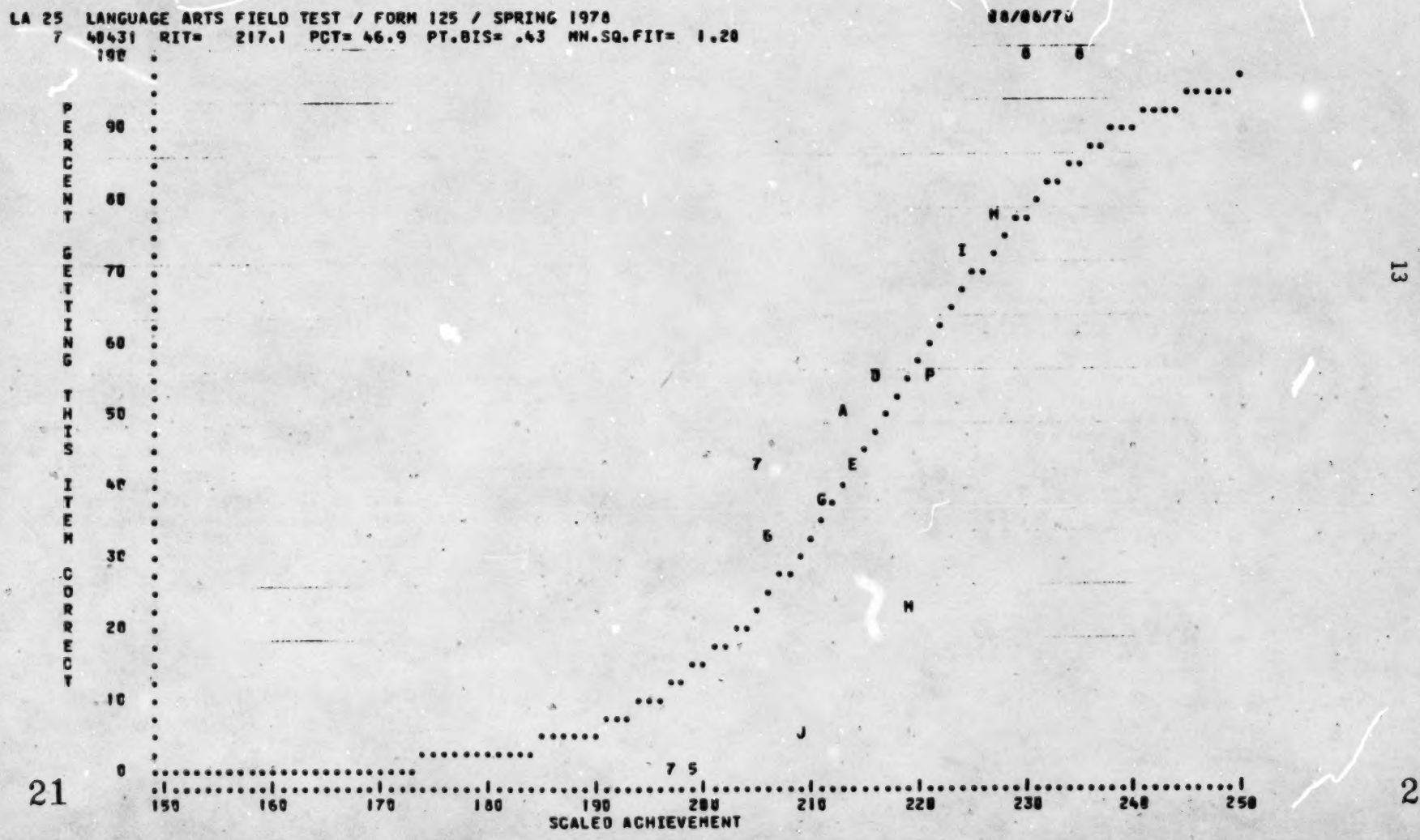

LA 25 LANGUAGE ARTS FIELD TEST / FORM 125 / SPRING 1978
7 40431 RIT= 217.1 PCT= 46.9 PT.BIS= .43 MN.SQ.FIT= 1.28
08/06/78
PERCENT GETTING THIS ITEM CORRECT
100
90
80
70
60
50
40
30
20
10
0
150
160
170
180
190
200
210
220
230
240
250
SCALED ACHIEVEMENT

Figure 7b

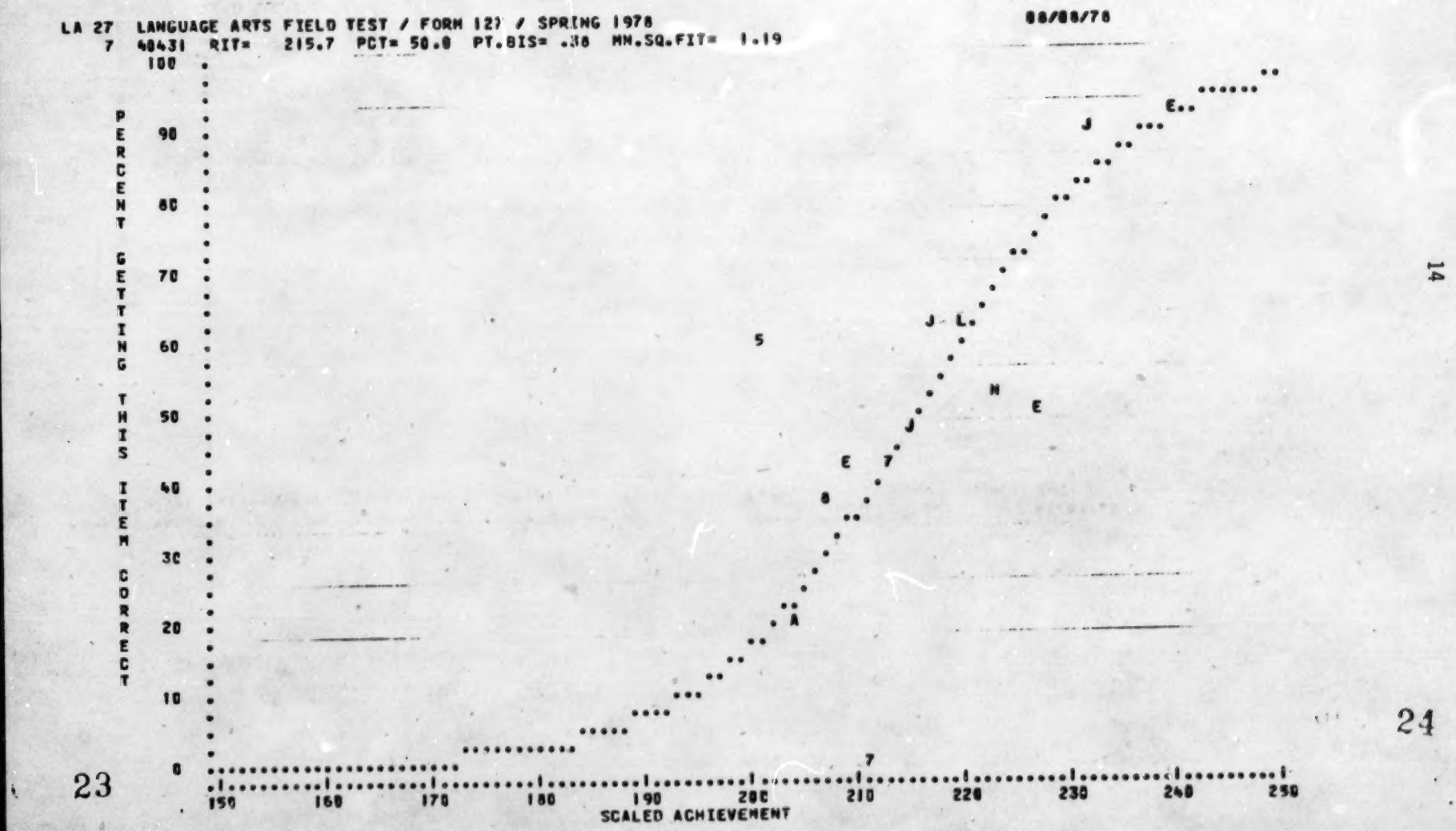
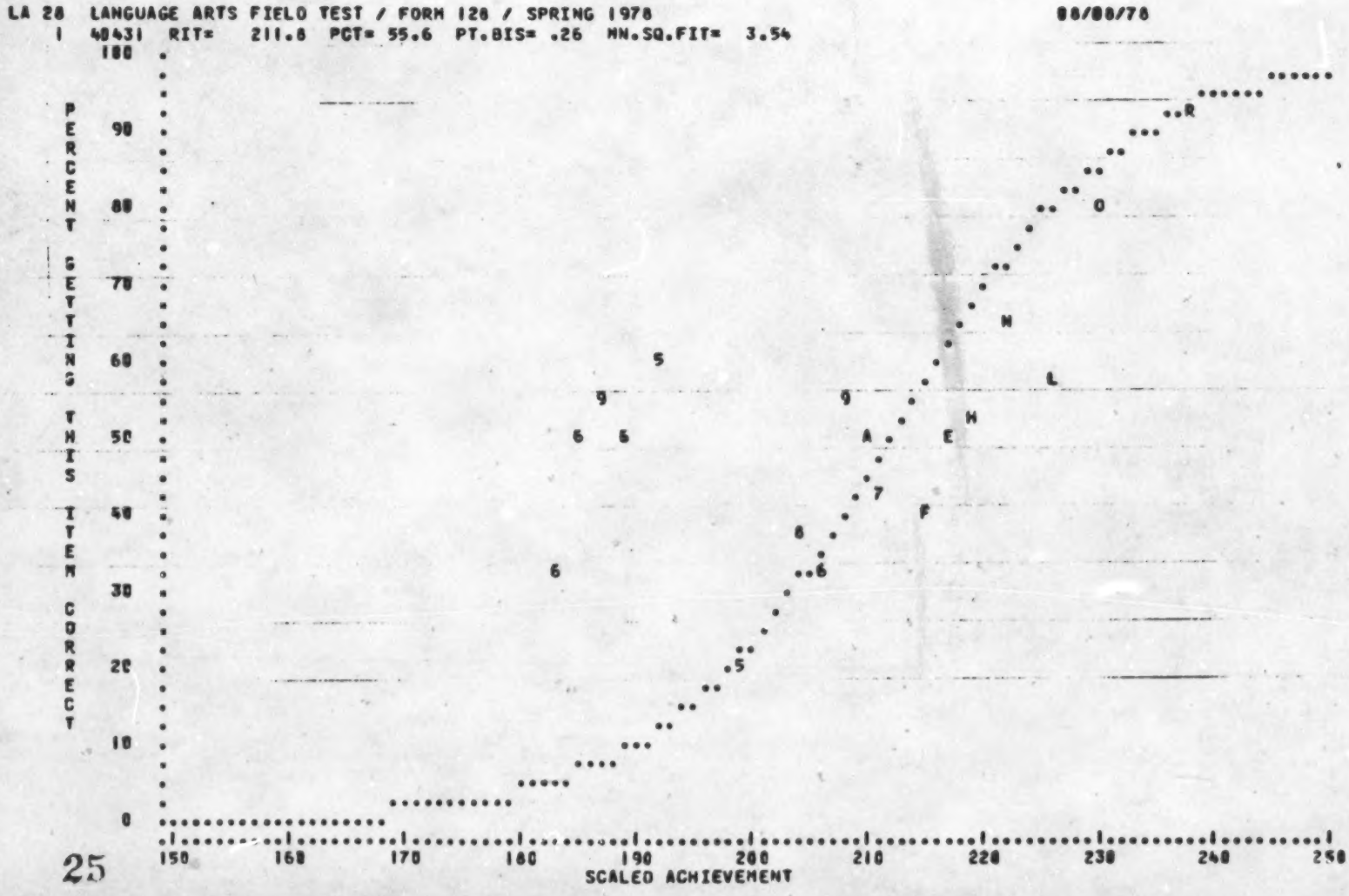

LA 27 LANGUAGE ARTS FIELD TEST / FORM 121 / SPRING 1978
7 40431 RIT= 215.7 PCT= 50.0 PT.BIS= .38 MN.SQ.FIT= 1.19
00/00/78
PERCENT GETTING THIS ITEM CORRECT
100
90
80
70
60
50
40
30
20
10
0
150
160
170
180
190
200
210
220
230
240
250
SCALED ACHIEVEMENT

Figure 7c


LA 28 LANGUAGE ARTS FIELD TEST / FORM 128 / SPRING 1978
I 40431 RIT= 211.8 PCT= 55.6 PT.BIS= .26 MN.SQ.FIT= 3.54
00/00/78
PERCENT GETTING THIS ITEM CORRECT
100
90
80
70
60
50
40
30
20
10
0
150
160
170
180
190
200
210
220
230
240
250
SCALED ACHIEVEMENT

DISCUSSION & CONCLUSIONS

The method described in this report for combining information across administrations of the same items in different tests to different student populations takes advantage of the fact that each item has an inherent level of difficulty independent of the group to which it is administered. Thus the performance of two or more groups on an item can be plotted in relation to the item's inherent level of difficulty and the resulting item characteristic curves can be shifted along the achievement baseline until the corresponding theoretical curves coincide. The result is a composite ICC which summarizes the information from all the administrations of the item.

The major findings from this study which included over 700 items administered to three student groups were:

1. When the ICC had a lower asymptote, it was close to or below the theoretical curve's asymptote for the items in these tests. Therefore, there is no evidence ~~that~~ of unusual resorting to guessing by students in the lowest achievement levels. It therefore appears inappropriate to include a guessing parameter in the scaling of the items used in a school testing program unless there is definite evidence of guessing.

2. The point biserial and mean square fit often gave conflicting information about the quality of the same item, and their values varied from administration to administration. Therefore, neither is a reliable indicator of item quality. At present, there appears to be no statistic or combination of statistics which can substitute for the item characteristic curve.

3. The method of combining information across administrations significantly increases the amount and utility of information available about the actual performance of an item.

APPENDIX A

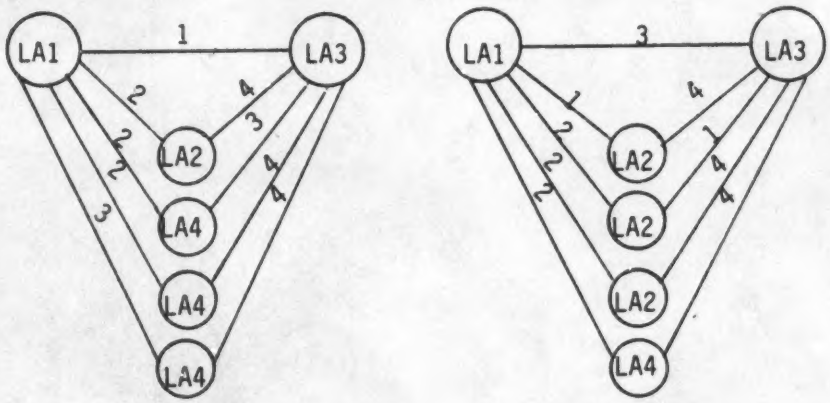
THE FOURSQUARE LINKING NETWORK

Fred Forster
George Ingebo
Portland Public Schools

This network was used to establish the NWEA language usage item bank from a large number of previously untried items. As shown in Figure 1, the Foursquare network is based on the principle of calibrating each item three times to produce two direct and four confirming values for each link, which makes it ideal for determining and excluding questionable data in building the bank. By numbering each block of ten items, the design for the field tests is shown in Table 1. The pattern includes 88 blocks of ten items each (880 items), arranged in 64 tests requiring approximately 13,000 students, giving a yield of 13.75 items per test.

As shown in Figure 2, this linking network derives its name from the fact that each group of four tests can be analyzed to give a pool of 80 items. Each group of four 80-item pools can then be analyzed to give a 240-item pool. The total of four 240-item pools can then be analyzed to give a single 880-item pool. Thus, there are three successive "phases" of the analysis leading to the final item pool.

In analyzing each of the links between two tests (say LA1 and LA3) the following triangulations can be made:



This configuration then provides two direct and four confirming values for each link. Once tests LA1 through LA4 are combined into a single pool,

Figure 1

Phase I of the Foursquare Linking Network

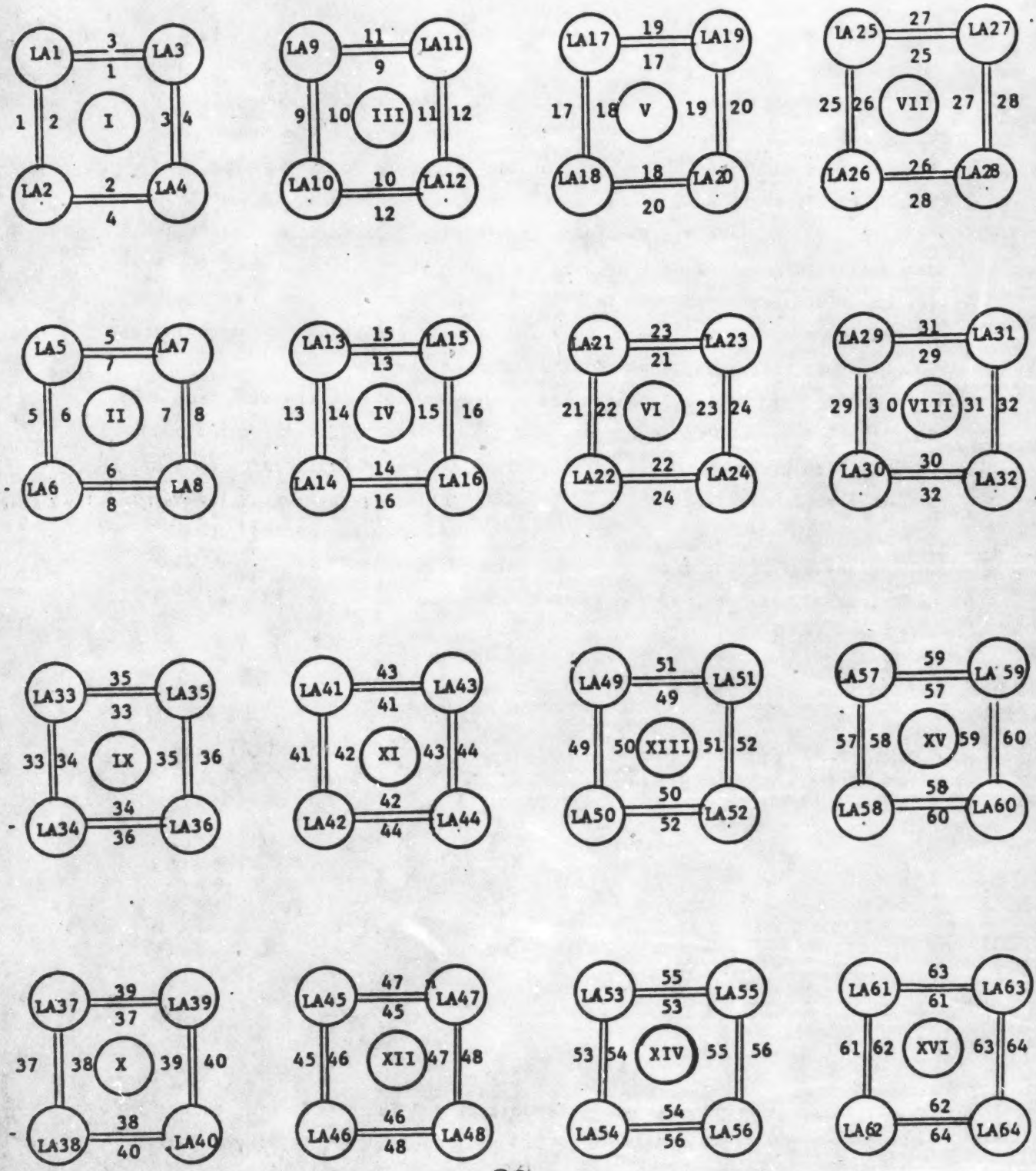


Table 1

The Foursquare Linking Pattern

			Phase I								Phase II				
TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III	TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III
LA1	I	A	1	2	3	65	x	LA17	V	B	17	18	19	69	x
LA2	I	A	1	2	4	66	x	LA18	V	B	17	18	20	70	x
LA3	I	A	1	3	4	67	x	LA19	V	B	17	19	20	71	x
LA4	I	A	2	3	4	x	81	LA20	V	B	18	19	20	x	81
LA5	II	A	5	6	7	65	x	LA21	VI	B	21	22	23	69	x
LA6	II	A	5	6	8	66	x	LA22	VI	B	21	22	24	70	x
LA7	II	A	5	7	8	68	x	LA23	VI	B	21	23	24	72	x
LA8	II	A	6	7	8	x	82	LA24	VI	B	22	23	24	x	82
LA9	III	A	9	10	11	65	x	LA25	VII	B	25	26	27	69	x
LA10	III	A	9	10	12	67	x	LA26	VII	B	25	26	28	71	x
LA11	III	A	9	11	12	68	x	LA27	VII	B	25	27	28	72	x
LA12	III	A	10	11	12	x	83	LA28	VII	B	26	27	28	x	84
LA13	IV	A	13	14	15	66	x	LA29	VIII	B	29	30	31	70	x
LA14	IV	A	13	14	16	67	x	LA30	VIII	B	29	30	32	71	x
LA15	IV	A	13	15	16	68	x	LA31	VIII	B	29	31	32	72	x
LA16	IV	A	14	15	16	x	*85	LA32	VIII	B	30	31	32	x	*86

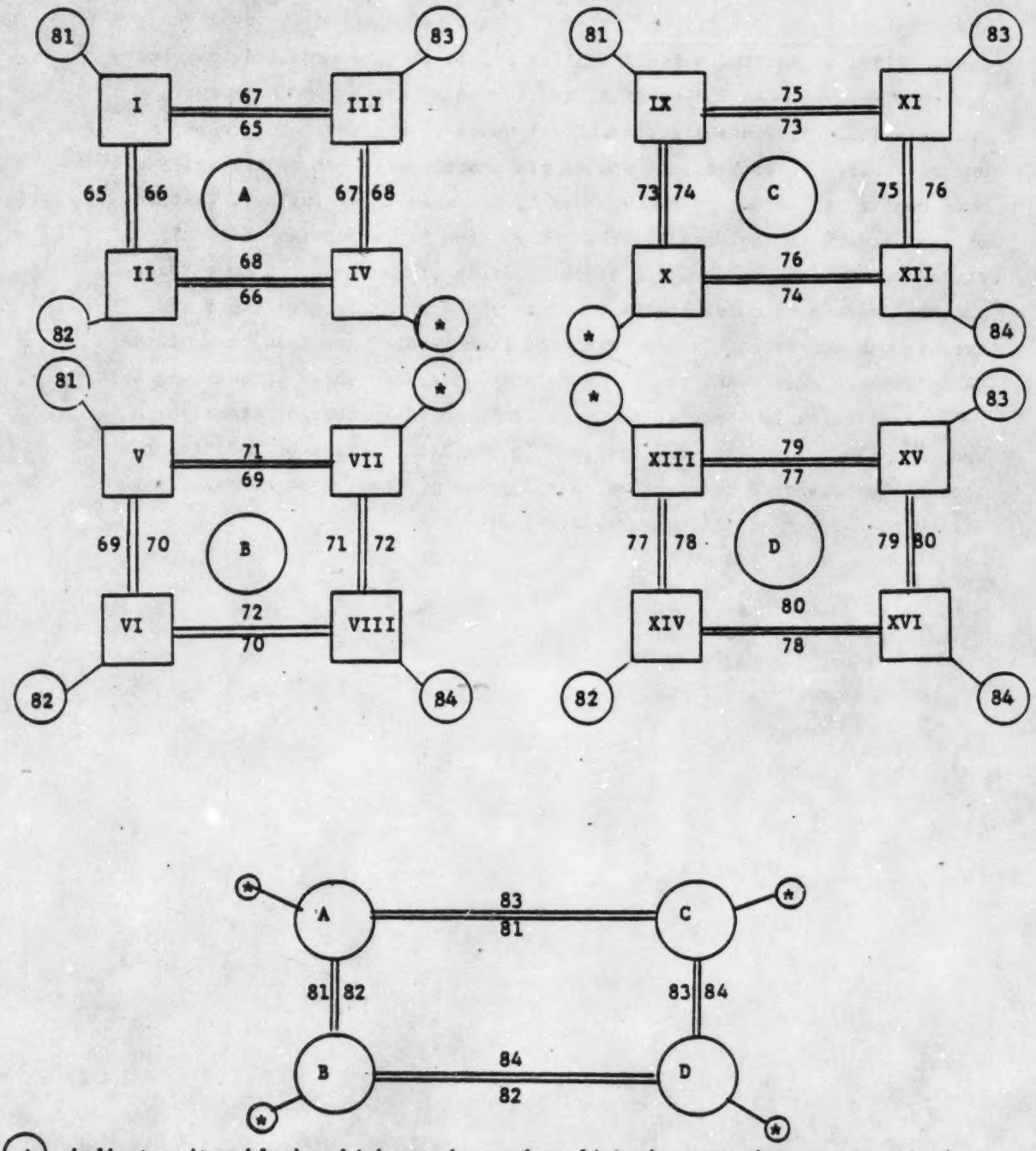
* Available for linking to higher levels.

			Phase I								Phase II				
TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III	TEST	PHASE II	PHASE III	1	2	3	PHASE II	PHASE III
LA33	IX	C	33	34	35	73	X	LA49	XIII	D	49	50	51	77	X
LA34	IX	C	23	34	36	74	X	LA50	XIII	D	49	50	52	78	X
LA35	IX	C	33	35	36	75	X	LA51	XIII	D	49	51	52	79	X
LA36	IX	C	34	35	36	X	81	LA52	XIII	D	50	51	52	X	82
LA37	X	C	37	38	39	73	X	LA53	XIV	D	53	54	55	77	X
LA38	X	C	37	38	40	74	X	LA54	XIV	D	53	54	56	78	X
LA39	X	C	37	39	40	76	X	LA55	XIV	D	53	55	56	80	X
LA40	X	C	38	39	40	X	83	LA56	XIV	D	54	55	56	X	83
LA41	XI	C	41	42	43	73	X	LA57	XV	D	57	58	59	77	X
LA42	XI	C	41	42	44	75	X	LA58	XV	D	57	58	60	78	X
LA43	XI	C	41	43	44	76	X	LA59	XV	D	57	59	60	80	X
LA44	XI	C	42	43	44	X	84	LA60	XV	D	58	59	60	X	84
LA45	XII	C	45	46	47	74	X	LA61	XVI	D	61	62	63	78	X
LA46	XII	C	45	46	48	75	X	LA62	XVI	D	61	62	64	79	X
LA47	XII	C	45	47	48	76	X	LA63	XVI	D	61	63	64	80	X
LA48	XII	C	46	47	48	X	*87	LA64	XVI	D	62	63	64	X	*88

*Available for linking to higher levels

Figure 2

Phase II of the Foursquare Linking Network



(*) indicates item blocks which may be used to link the network to an item bank or another network.

blocks 65, 66, and 67 link to the four adjacent pools (each formed by a group of four tests) through a comparable analysis.

The redundancy of direct links in this network strengthens the design and simplifies the identification of bad test information in any given test. Since organizing a field testing effort of this magnitude may leave little time available for readministration of a test shown to be weak, the redundancy provides a second direct value if one turns out to be unacceptable. If both direct values are unacceptable, there are still six more confirming values to salvage the link. When these failsafe features of the network are not needed, the two 10 item links between each pair of tests are combined to form six strong 20 item links among the four tests. Another feature of this network is that when sixteen or more tests are used it can accommodate a wide range of item levels. (In the case of the NWEA language arts bank, the 64 tests spanned grades three through eight.)

Theoretically there is no upper limit to the number of items which can be calibrated using this design. In addition, it can be modified to produce "unbalanced" networks so that a group of four tests can easily be linked to a group of sixteen or sixty-four.